L Data presentation and interpretation

L1

Interpret diagrams for single-variable data, including understanding that area in a histogram represents frequency.

Connect to probability distributions.

Students should be able to:

- interpret box and whisker plots (boxplots), cumulative frequency curves and histograms

  Note: students will **not** be expected to construct these diagrams.
- Comment on the skewness of a distribution shown in a boxplot. A distribution is positively skewed if Q3 – Q2 > Q2 – Q1 and negatively skewed if Q3 – Q2 < Q2 – Q1
- use diagrams to find probabilities of given events.
  - Note: when studying this topic, students can use data from the large data set and process it using software such as GeoGebra (geogebra.org)
- interpret unfamiliar graphs or representations of data.

L Data presentation and interpretation

L4

Recognise and interpret possible outliers in data sets and statistical diagrams.

Select or critique data presentation techniques in the context of a statistical problem.

Be able to clean data, including dealing with missing data, errors and outliers.

Students should be able to:

- identify outliers either from a given rule or from observation of a given diagram
- comment on the likely effect of removing the outlier
- identify clear errors in data and comment on or suggest subsequent actions needed
- select which of the representations in sections L1 and L2 is appropriate for representing given data sets
- criticise, in context, a given representation.

9.3 Single Variable Data

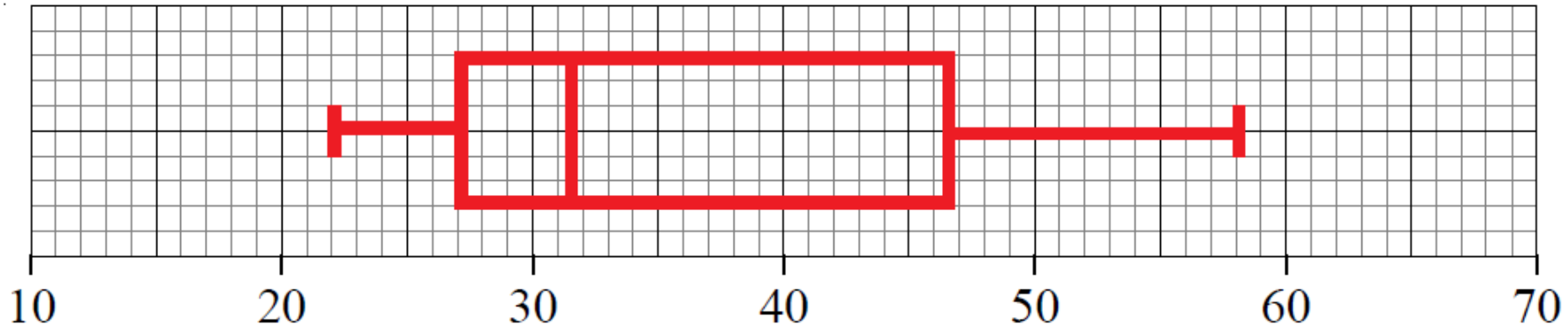
Box Plots – Basics from GCSE

1) The ages of 20 teachers are listed below.

27 31.5 46.5

22, 22, 24, 25, 27, 27, 28, 29, 29, 29, 34, 35, 41, 43, 44, 49, 55, 57, 58, 58

a) On the grid below, draw a box plot to show the information about the teachers.



b) What is the interquartile range of the ages of the teachers? 19.5 years

9.3 Single Variable Data

Box Plots

One quarter of the data values in the sample lie between each consecutive pair of vertical lines on the diagram.

Lines placed further apart show a greater spread of data.

Outliers are displayed on a box plot as crosses, they are not included in whiskers.

If you have sufficient information you should use the most extreme value that is not an outlier as the end of the whisker. Otherwise,

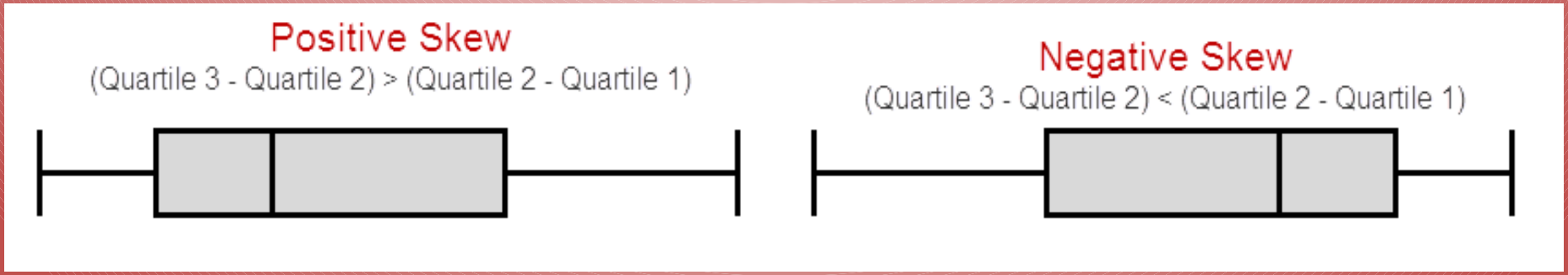
9.3 Single Variable Data

Box Plots

A distribution shown in a box plot can be positively or negatively **skewed**.

It is **positively** skewed when .

It is **negatively** skewed when .

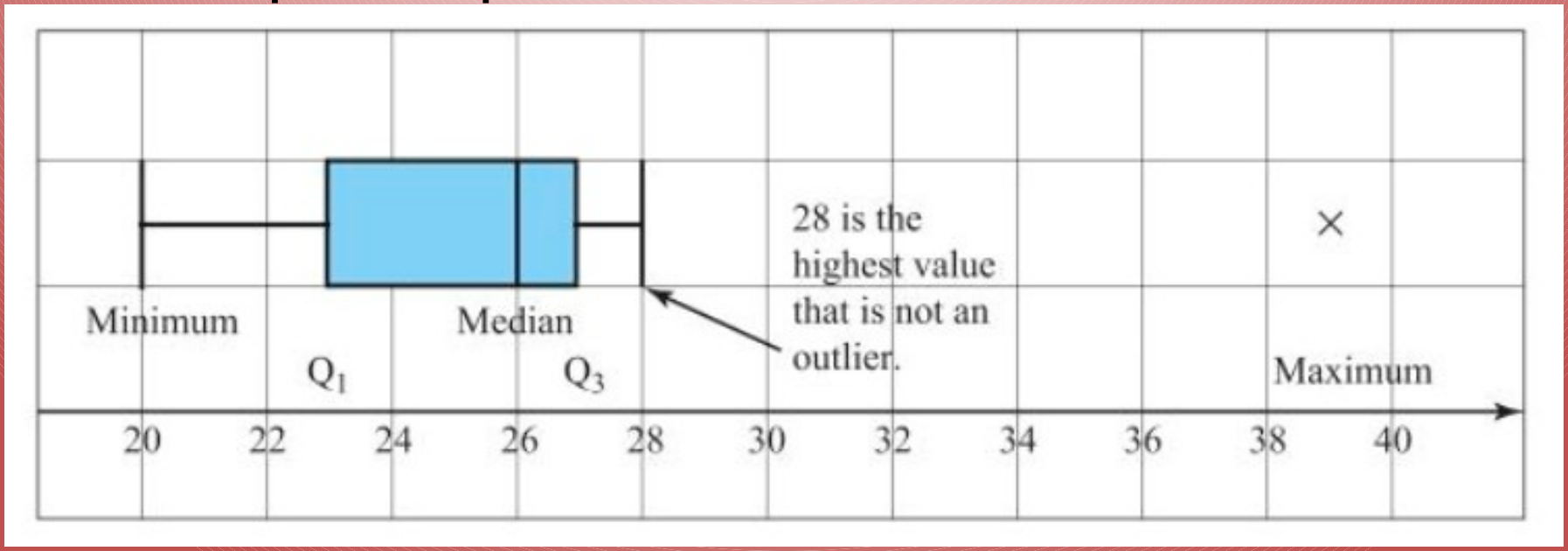


9.3 Single Variable Data

Box Plots

$$Q_1 - 1.5 \times IQR = 23 - (1.5 \times 4) = 17$$
$$Q_3 + 1.5 \times IQR = 27 + (1.5 \times 4) = 33$$

Suppose you define an outlier as a value less than
Q_1 – 1.5 x IQR or more than Q_3 + 1.5 x IQR.
This boxplot represents the set of data:



9.3 Single Variable Data

Box Plots

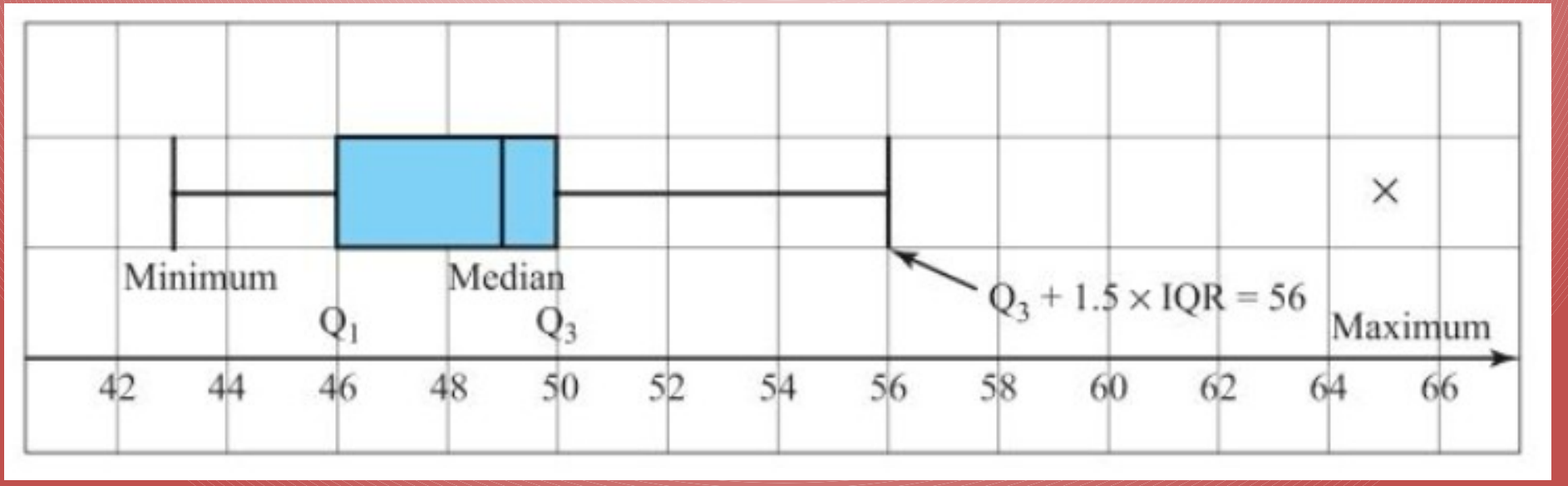
$Q_1 - 1.5 \times IQR = 46 - (1.5 \times 4) = 40$

$Q_3 + 1.5 \times IQR = 50 + (1.5 \times 4) = 56$

Suppose you define an outlier as a value less than

Q_1 – 1.5 x IQR or more than Q_3 + 1.5 x IQR.

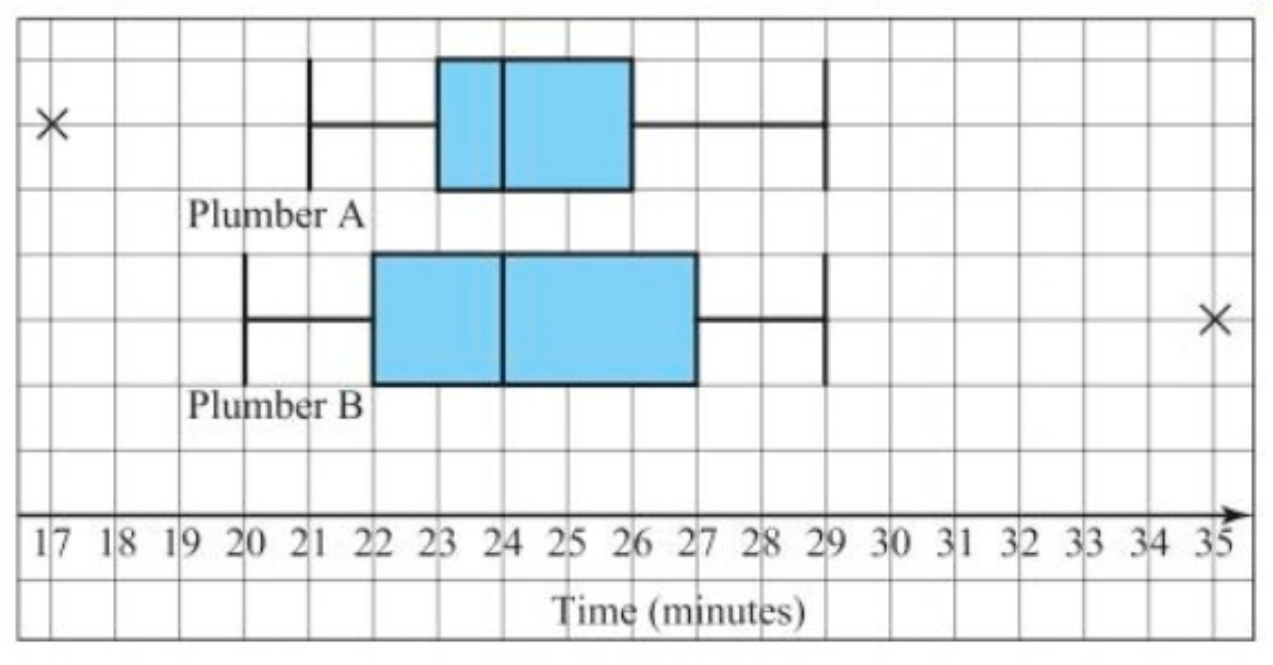
This boxplot represents the set of data



9.3 Single Variable Data

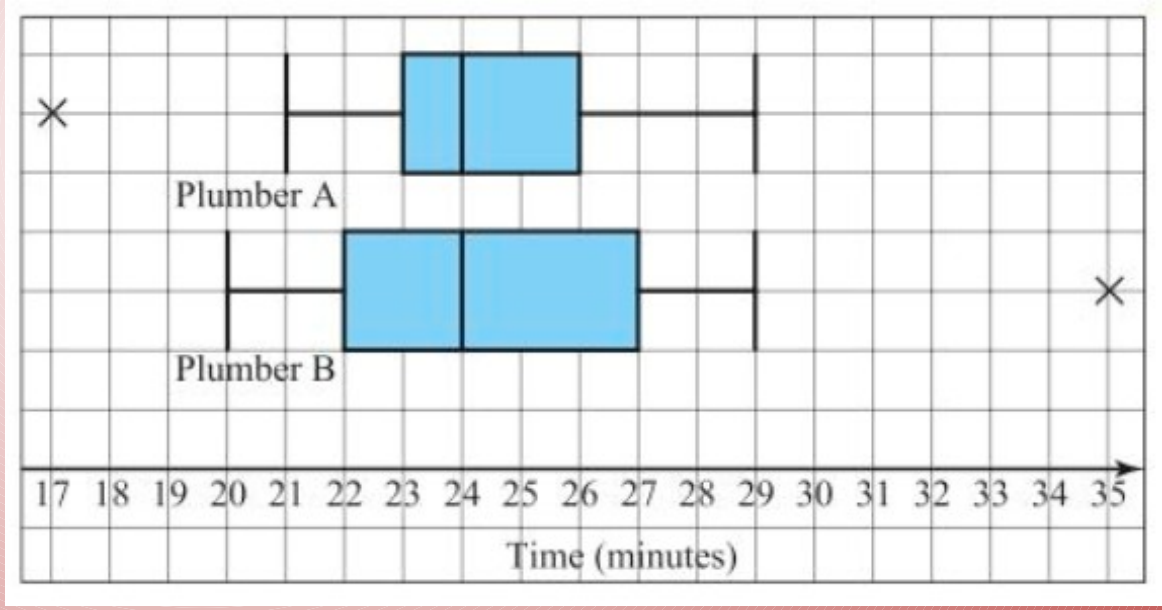
Box Plots – Example 1

A building company works with two plumbers. Over a period of time, they assess how long it takes each plumber to fix leaking pipes. This data is



9.3 Single Variable Data

Box Plots – Example 2



An outlier is defined as a value less than Q_1 – 1.5 x IQR or more than Q_3 + 1.5 x IQR.

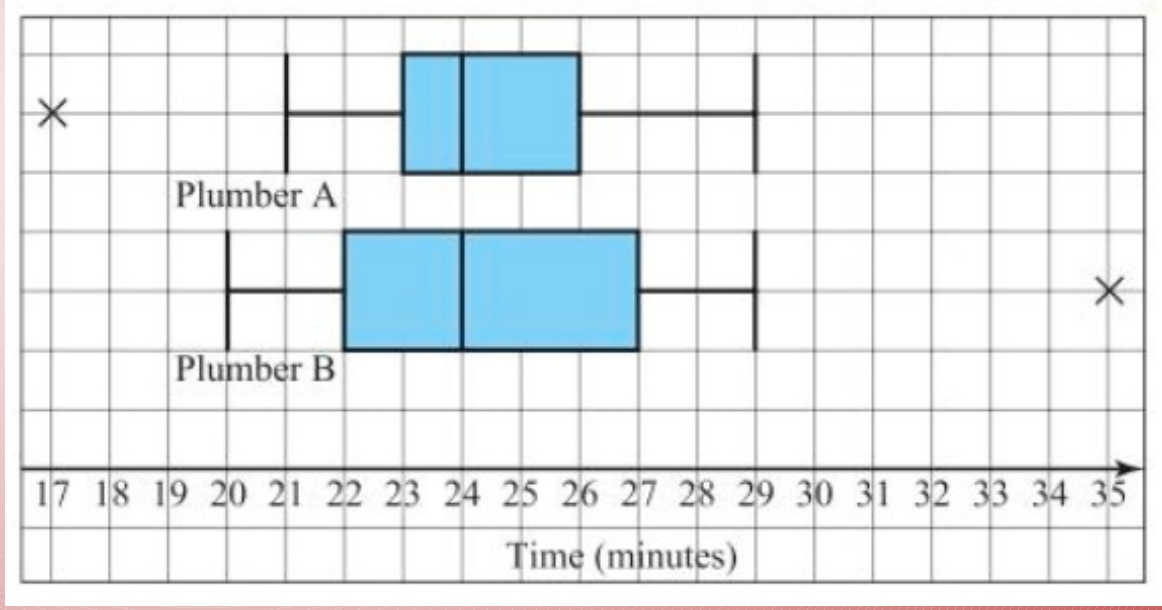
a) Write down the minimum, lower quartile, median, upper quartile and

A: min = 17, Q_1 = 23, median = 24, Q_3 = 26, max = 29

B: min = 20, Q_1 = 22,

9.3 Single Variable Data

Box Plots – Example 2



b) Recommend a choice of plumber given that no outliers are deleted.

It would be most sensible to choose plumber A. Both have a median of 24 minutes however plumber A's data shows less variation. It has a smaller IQR of 3 compared to 5 for plumber B. Plumber A has an outlier representing a quick time, plumber B's outlier represents a slower

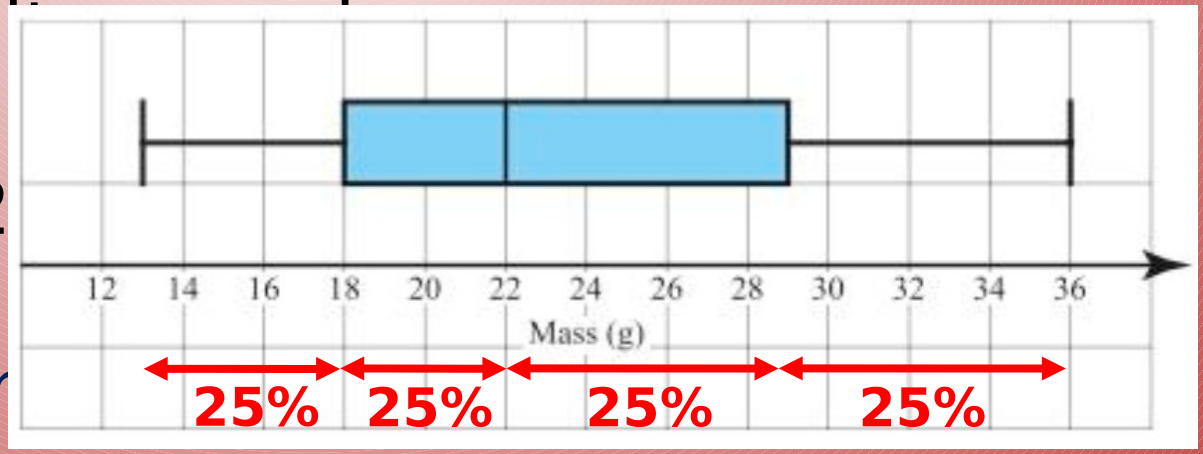
9.3 Single Variable Data

You can estimate probabilities of certain events happening by using diagrams that show grouped data. This often involves assuming data is equally distributed.

9.3 Single Variable Data

Example 3

This box plot shows the masses (in grams) of 52 eggs from a certain species of bird. An egg is picked at random from this set. Estimate the probability that it



a) More than 22

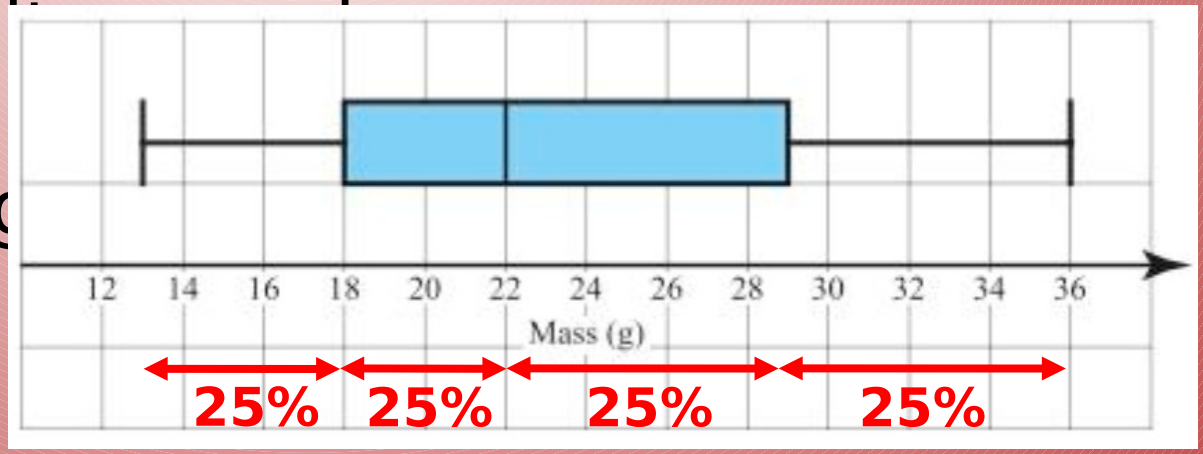
22 is the median, so the probability is 50%

9.3 Single Variable Data

Example 3

This box plot shows the masses (in grams) of 52 eggs from a certain species of bird. An egg is picked at random from this set. Estimate the probability that

b) Less than 18g

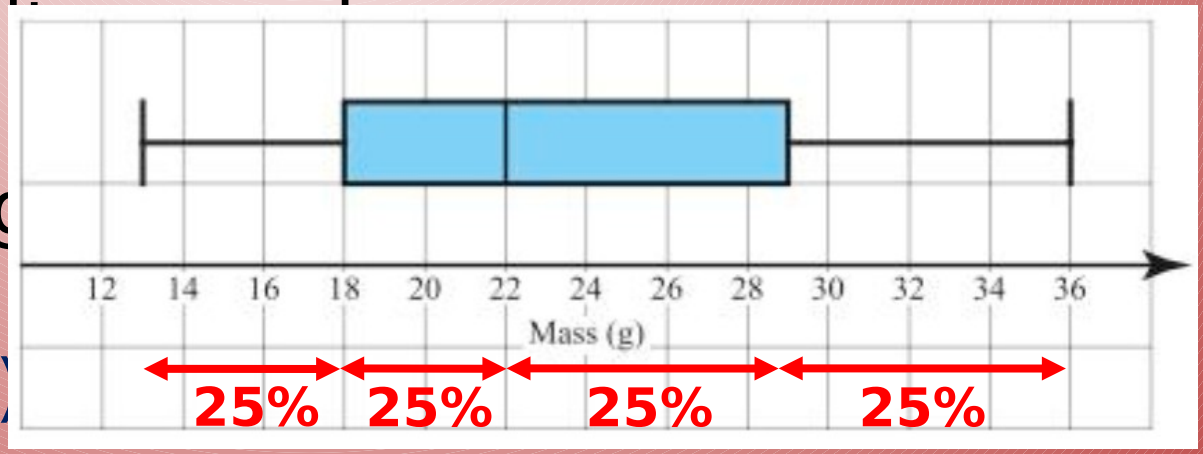


18 is the lower quartile so the probability is 25%

9.3 Single Variable Data

Example 3

This box plot shows the masses (in grams) of 52 eggs from a certain species of bird. An egg is picked at random from this set. Estimate the probability that



c) Less than 20g

25% + 0.5(25%)
= 37.5%

9.3 Single Variable Data

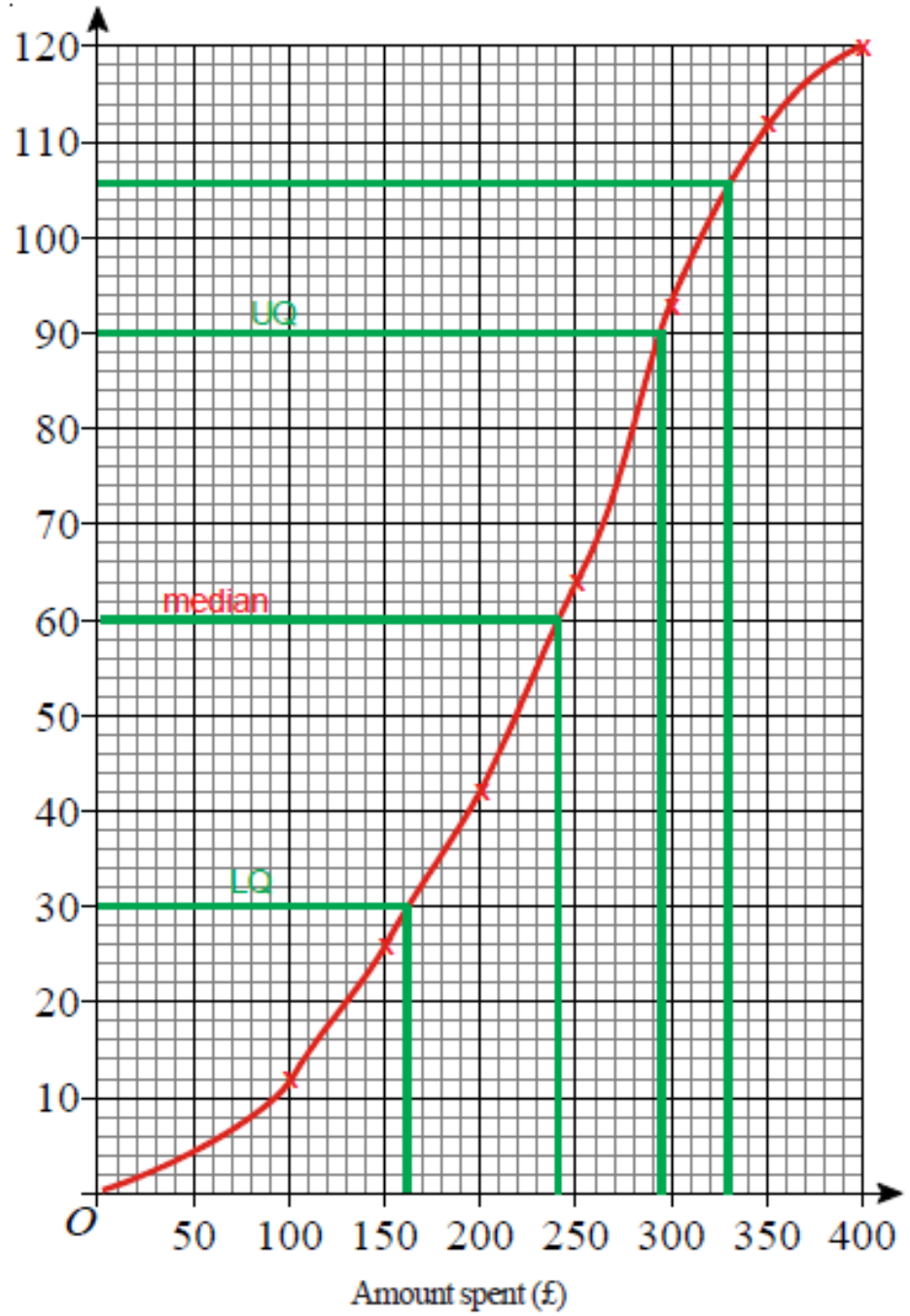
Cumulative Frequency – Basics from GCSE

Fred did a survey about the amount of money spent by 120 men at Christmas.

The cumulative frequency table gives some information about the amounts of money spent by the 120 men.

Amount (£A) spent	Cumulative frequency
$0 < A \leqslant 100$	12
$0 < A \leqslant 150$	26
$0 < A \leqslant 200$	42
$0 < A \leqslant 250$	64
$0 < A \leqslant 300$	93
$0 < A \leqslant 350$	112
$0 < A \leqslant 400$	120

a) On the grid, draw a cumulative frequency diagram.



b) Use your cumulative frequency diagram to estimate the median.

£240

c) Use your cumulative frequency diagram to estimate the interquartile range of the amount of money spent. £295 – £160 = £135

d) Use your cumulative frequency diagram to estimate the number of men who spent more than £330.

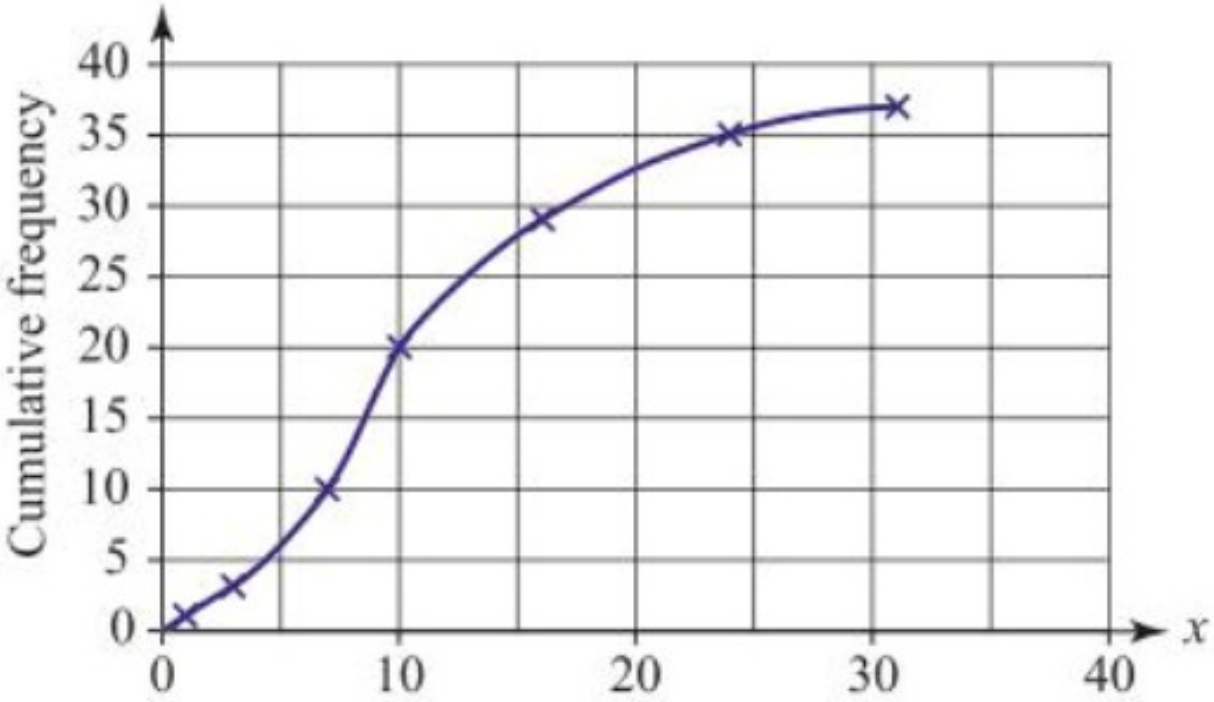
14

Amount (£A) spent	Cumulative frequency
$0 < A \leqslant 100$	12
$0 < A \leqslant 150$	26
$0 < A \leqslant 200$	42
$0 < A \leqslant 250$	64
$0 < A \leqslant 300$	93
$0 < A \leqslant 350$	112
$0 < A \leqslant 400$	120

Example 4

The heights of a sample of a species of plant are recorded. Complete the frequency table and use the cumulative frequency graph to fill in the

Height in cm	$0 \le x < 1$	$1 \le x < 3$	$3 \le x < 7$	$7 \le x < 10$	$10 \le x < 16$	$16 \le x < 24$	$24 \le x < 31$	$31 \le x$
Frequency f	1	2			9			0



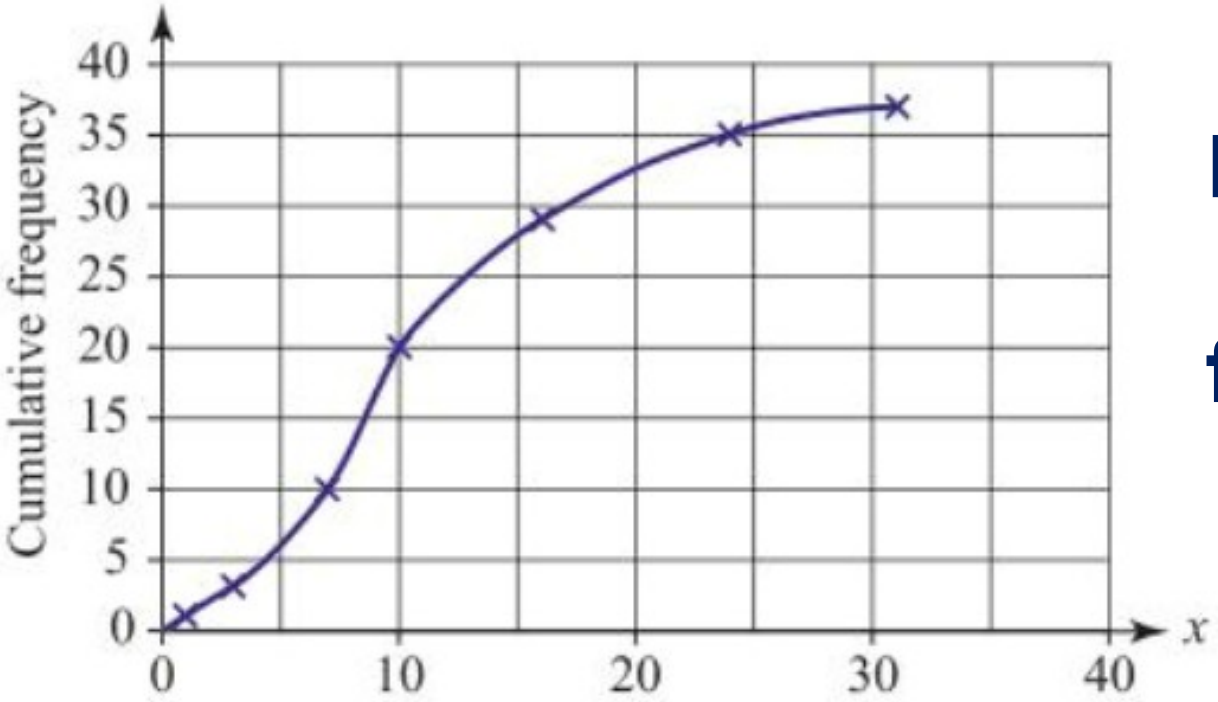
For 3 ≤ x < 7,

f = 10 – 3 = 7

Example 4

The heights of a sample of a species of plant are recorded. Complete the frequency table and use the cumulative frequency graph to fill in the

Height in cm	$0 \le x < 1$	$1 \le x < 3$	$3 \le x < 7$	$7 \le x < 10$	$10 \le x < 16$	$16 \le x < 24$	$24 \le x < 31$	$31 \le x$
Frequency f	1	2	**7**		9			0



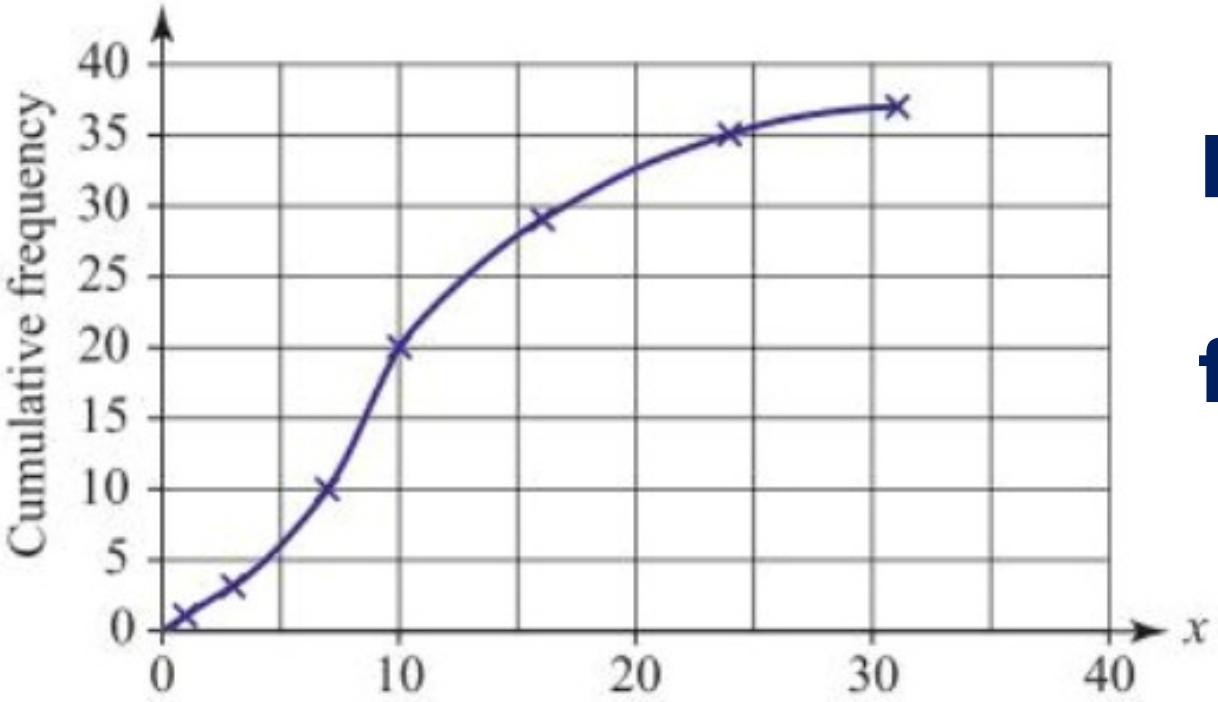
For 7 ≤ x < 10

f = 20 – 10 = 1

Example 4

The heights of a sample of a species of plant are recorded. Complete the frequency table and use the cumulative frequency graph to fill in the

Height in cm	$0 \le x < 1$	$1 \le x < 3$	$3 \le x < 7$	$7 \le x < 10$	$10 \le x < 16$	$16 \le x < 24$	$24 \le x < 31$	$31 \le x$
Frequency f	1	2	**7**	**10**	9			0



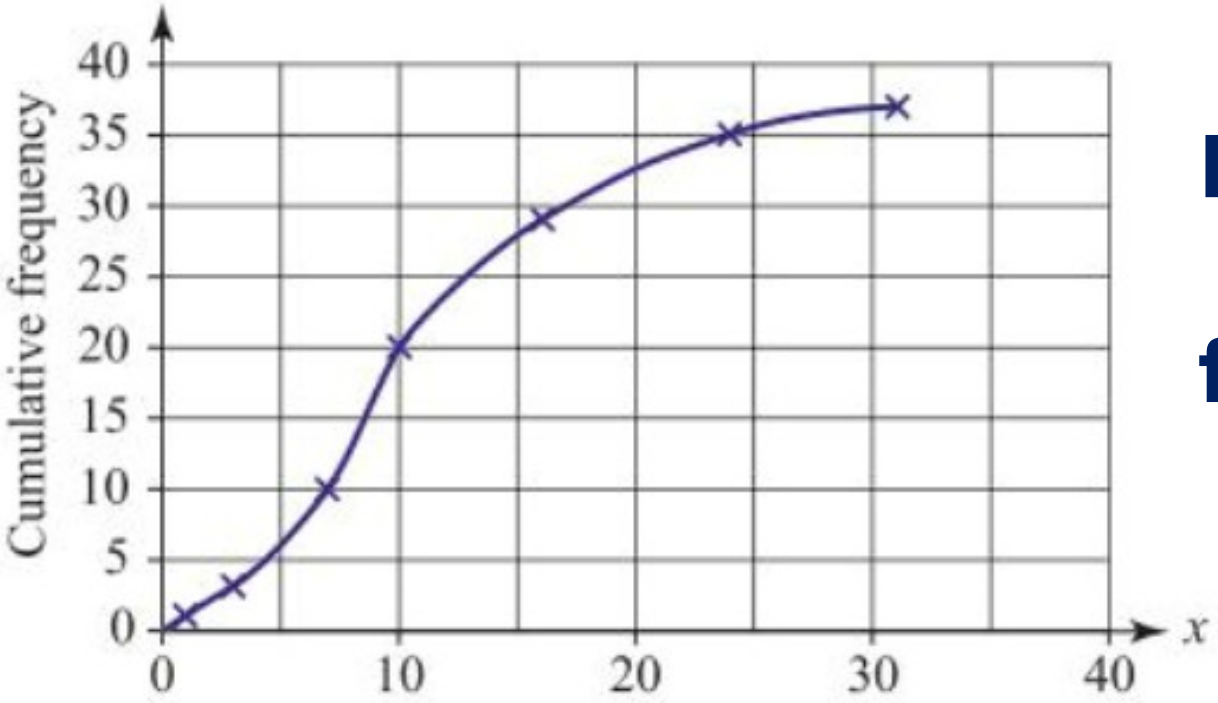
For 16 ≤ x < 2

f = 35 – 29 = 6

Example 4

The heights of a sample of a species of plant are recorded. Complete the frequency table and use the cumulative frequency graph to fill in the

Height in cm	$0 \le x < 1$	$1 \le x < 3$	$3 \le x < 7$	$7 \le x < 10$	$10 \le x < 16$	$16 \le x < 24$	$24 \le x < 31$	$31 \le x$
Frequency f	1	2	**7**	**10**	9	**6**	**2**	0



For 24 ≤ x < 3

f = 37 – 35 = 2

9.3 Single Variable Data

Histograms

Histograms show the **distribution** (shape) of the data.

The distribution of data is how often each outcome occurs. Each outcome occurs with a given **frequency**.

For grouped data, the groups must be consecutive, non-overlapping ranges and do not have to be equal in width.

9.3 Single Variable Data

Histograms

There are no gaps between the bars.

The height of the bar often represents **frequency density**.

The width of the bar is the size of the interval.

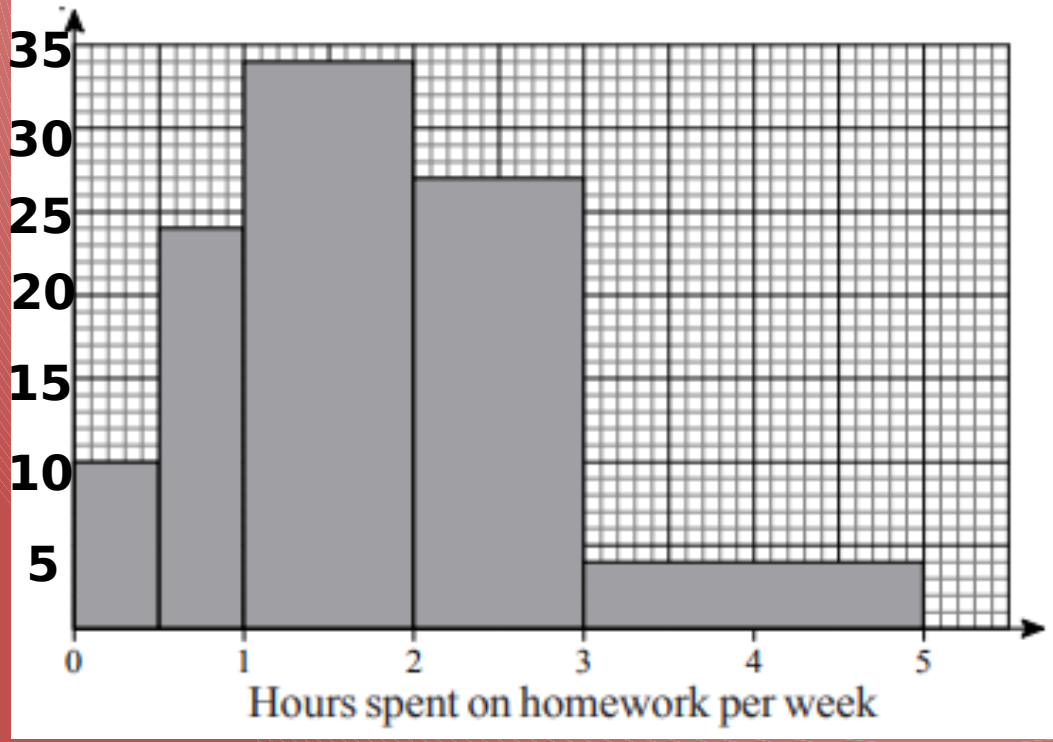
Frequency is **proportional** to the area of each bar.

You will not be expected to draw a histogram. You will be expected to find probabilities or find the mean/standard deviation

9.3 Single Variable Data

Histograms – Example 5

The histogram shows the amount of time, in hours, that students spend on their homework per week.



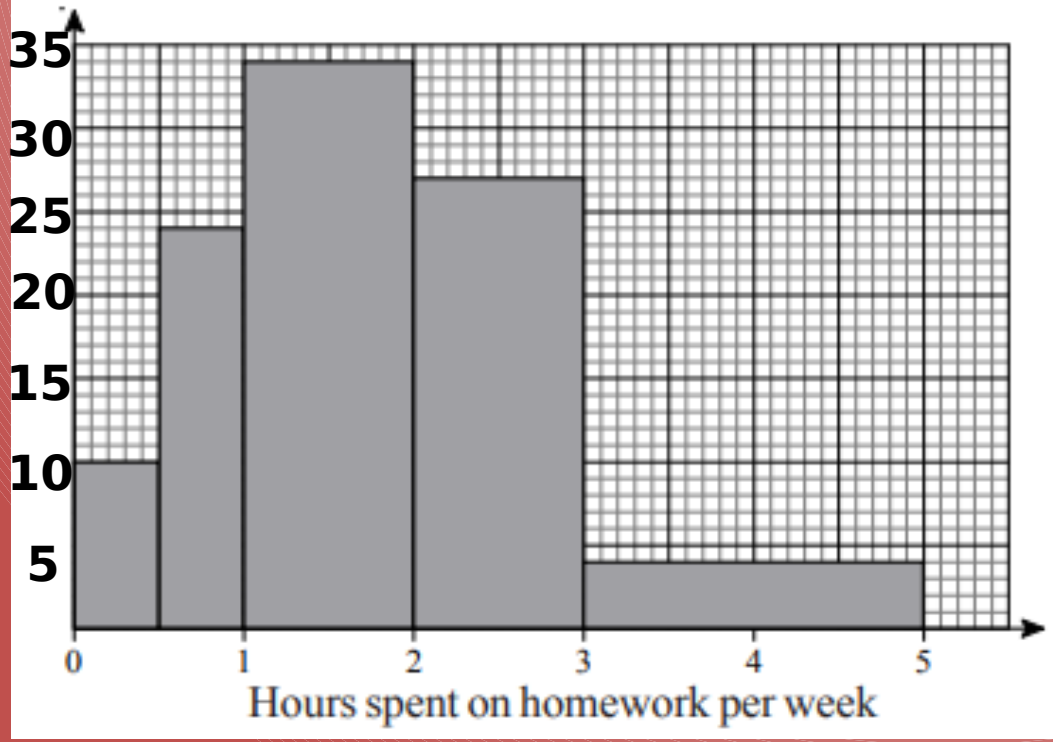
Find the mean and standard

Hours	Frequency
0 – 0.5	0.5 x 10 = **5**
0.5 – 1	0.5 x 24 = **12**
1 – 2	1 x 34 = **34**
2 – 3	1 x 27 = **27**

9.3 Single Variable Data

Histograms – Example 5

The histogram shows the amount of time, in hours, that students spend on their homework per week.



Find the mean and standard

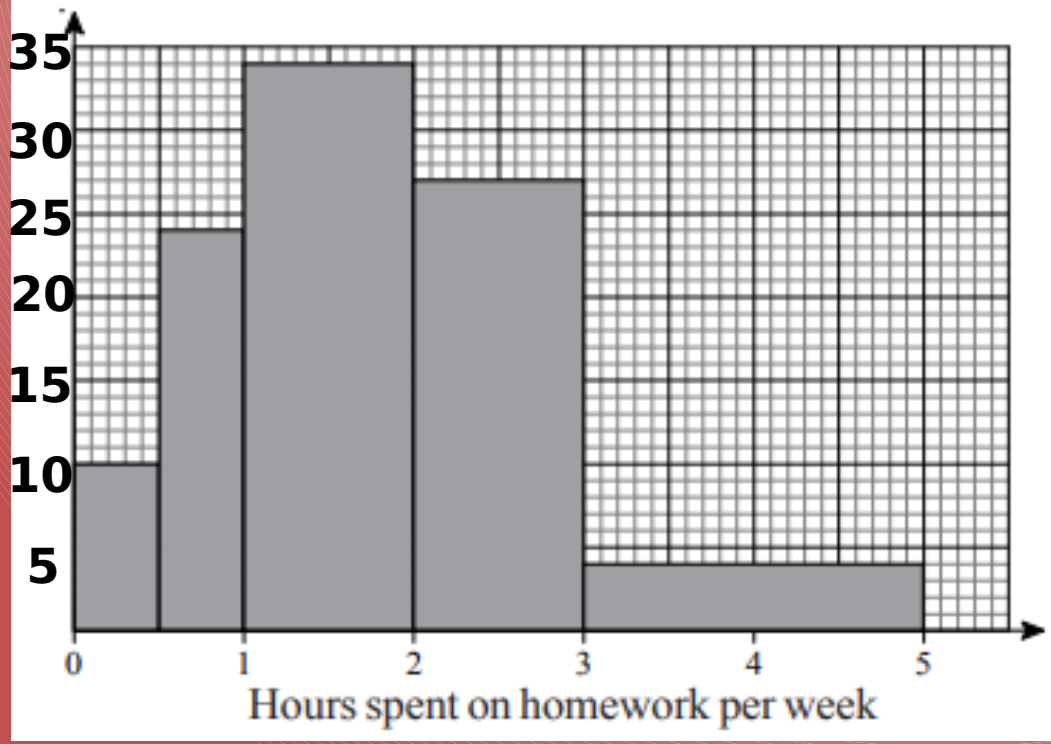
Hours	Frequency
0.25	0.5 x 10 = **5**
0.75	0.5 x 24 = **12**
1.5	1 x 34 = **34**
2.5	1 x 27 = **27**

From calc:

9.3 Single Variable Data

Histograms – Example 5

The histogram shows the amount of time, in hours, that students spend on their homework per week.



Hours	Frequency
0 – 0.5	0.5 x 10 = **5**
0.5 – 1	0.5 x 24 = **12**
1 – 2	1 x 34 = **34**
2 – 3	1 x 27 = **27**
3 – 5	2 x 4 = **8**

Find the probability that a student chosen at random spends more than 2 hours a week on their homework.

$\frac{35}{86}$

9.3 Single Variable Data

	Advantages	Disadvantages
Box Plot	Highlights outliers. Makes it easy to compare data sets.	Data is grouped into only four categories so some details analysis is not possible.
Histogram	Clearly shows shape of distribution.	Doesn't always highlight outliers. It is possible but not easy to estimate Q_1, Q_2 and Q_3.
Cumulative Frequency Curve	Makes it easy to find the five number summary.	Doesn't always highlight outliers. If interval boundaries are not shown the degree of detail is not clear.

9.3 Single Variable Data

Example 6

A dietitian records the mean consumption in ml of dairy desserts (not frozen) per week for a group of clients on the same diet programme:
40, 41, 41, 44, 48, 51, 53, 53, 54, 54, , 59, 61, 62, 62, 63, 64, 65, 65, 66, 66, , 90

and are unknown values but the list is known to be in ascending order. An outlier is defined as a value less than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$

a) Explain why a box plot is not an appropriate choice of diagram to represent this data

9.3 Single Variable Data

Example 6

40, 41, 41, 44, 48, **51**, 53, 53, 54, 54, a, 59, 61, 62, 62, 63, 64, **65**, 65, 66, 66, b, 90

a) Explain why a box plot is not an appropriate choice of diagram to represent this data

6^{th} value 18^{th} value

90 is known to be an outlier. Hence b may or may not be an outlier so it would be impossible to plot the upper limit of the data accurately.

9.3 Single Variable Data

Example 6

40, 41, 41, 44, 48, **51**, 53, 53, 54, 54, a, 59, 61, 62, 62, 63, 64, **65**, 65, 66, 66, b, 90

b) Describe a more appropriate diagram to represent this data

A histogram would be better. As the relative values of and are known in relation to values either side, they would not affect the shape of the histogram as long as they are not used as boundaries of categories.

Dividing the data into several groups would